STAT 452 Project 2

**Section 1: ROC curves**

* Wish to display *specificity* and *sensitivity* of a classification model (for example, LDA)
* *Specificity*: percentage of non-defaulters that are correctly identified (true negative rate)
* *Sensitivity*: percentage of true defaulters that are identified (true positive rate)
* ROC curve plots 1-specificity (false positive rate) on the x-axis, and sensitivity on the y-axis. Ideally, we want a high sensitivity and a low 1-specificity
* ROC curve displays the 2 errors over all possible thresholds
* Default classifier (threshold) is 0.5, but can be higher/lower depending on the trade-off of the 2 errors
  + A lower threshold increases both 1-specificity and sensitivity, while a higher threshold decreases both
* Summary table:

|  |  |  |
| --- | --- | --- |
| **specificity** | True Negative Rate |  |
| **1-specificity** (x-axis) | False Positive Rate |  |
| **Sensitivity** (y-axis) | True Positive Rate |  |

* The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the (ROC) curve (**AUC**)
* The larger the AUC, the better the classifier (an ideal ROC curve will hug the top left corner)
* AUC should not be lower than 0.5, because at 0.5 it is the same as a classifier purely predicting by chance

**Section 2: Support Vector Machines (SVM)**

* Aims for better classification of most of the training observations, when the two classes cannot be separated by a hyperplane
* Allows for some observations to be on the wrong side of the margin/hyperplane for a more robust classification overall
* Maximizes the width of the margin M from the hyperplane, based on *slack variables* εi, and tuning parameter C such that
* If εi = 0 then the ith observation is on the correct side of the margin; εi > 0 then the ith observation is on the wrong side of the margin; εi > 1 then it is on the wrong side of the hyperplane
* C is the tolerance parameter, or the severity of the violations of the margin
* Observations that lie directly on the margin, or on the wrong side of the margin for their class, are known as *support vectors*, and only those vectors affect the classifier
* SVM is an extension of the support vector classifier such that it enlarges the feature space, using *kernels*, to accommodate non-linear boundary between classes
* Advantages:
  + Can handle non-linear relationships with kernels (logistic regression cannot)
  + Does not recursively split data (as in decision trees)
  + Robust to overfitting
* Disadvantages:
  + Sensitive to outliers (random forest is less sensitive)
  + Computationally expensive
  + Weak intuition